The Case Survey and Alternative Methods for Research Aggregation

Rand Corp Washington D C

Jun 74

THE CASE SURVEY AND ALTERNATIVE METHODS FOR RESEARCH AGGREGATION

William A. Lucas

June 1974

P-5252

The Rand Corporation
Santa Monica, California 90406

THE CASE SURVEY AND ALTERNATIVE METHODS FOR RESEARCH AGGREGATION*

William A. Lucas

The Rand Corporation, Washington, D.C.

A central argument used in defending basic scientific inquiry is that one does not have to prove the value of any one research project because it fits into a broader process of knowledge acquisition. As the knowledge base grows, it will cumulate and patterns will emerge that will provide a broader understanding of social life. Without that rationale, the burden of proof on each research project to prove its value becomes much more severe.

Yet we know the aggregative function in social research is poorly served. The profusion of publications and professional meetings, the sheer number of social scientists active in research, make it impossible to keep up with a broad literature. Psychology is probably representative when only half of the research "in 'core' journals will be read [or skimmed] by 1 percent or less of a random sample of psychologists."** The fragmentation of social science into disciplines makes the problem of keeping current easier, while creating artificial barriers between, for example, sociologists and political scientists doing very similar work. Different methodological orientations and correspondingly different journals further divide social scientists within disciplines. And the greatest barrier is between government contract research and the academic community. Yet, in the words of Robert Merton:

> But, for science to be advanced, it is not enough that fruitful ideas be originated or new experiments developed or new

---

*This paper was prepared for the Conference on Design and Measurement Standards in Political Science, Delavan, Wisconsin, May 1974. It draws heavily upon a line of aggregative research conducted by The Rand Corporation that began in 1972. Several substantive and methodological Rand reports are available and are referenced herein, but the author must acknowledge, in particular, the invaluable contributions and criticisms of Robert K. Yin.

**Reported in Robert K. Merton, "The Matthew Effect in Science," *Science*, Vol. 59, 1968, pp. 56-63.

> problems formulated or new methods instituted. The innovation must be effectively communicated to others . . . . For the development of science, only work that is effectively perceived and utilized by other scientists, then and there, matters.*

The purpose of this paper is to suggest that the cumulative function of research should itself be elevated to a central place in the profession and that we demand of it the same standards of scientific rigor we ask of any other type of inquiry. We shall begin by discussing three alternative approaches that can be used for aggregating research, and their strengths and weaknesses. Then the rules to be used in guiding aggregation will be treated. The emphasis will be placed on the third type of approach, case survey methods.

ALTERNATIVE APPROACHES TO AGGREGATIVE RESEARCH

There are at least three generic approaches to the task of aggregating research. The first and most common is the *propositional* approach: that of collecting statements of relationship from a set of studies. The other two are rarely employed, and their shared feature is that they use earlier research as sources of data, rather than as sources of conclusions or propositions. The *cluster* method involves the pooling and analysis of the original individual data input of the research being reviewed. The *case survey* approach relies on the descriptive materials in case studies of organizations, cultures, or some other common social unit. The information in the research reports becomes a form of data. No one approach or variation is always the best, and they can in fact be used in combination to strengthen the aggregation.

The Propositional Approach

In a propositional review, the conclusions of many studies are put in the form of statements of interrelationships among events, resources, processes, and outcomes. When there is agreement, the reviewer has a straightforward task. If similar studies have reached contradictory conclusions, the reviewer will seek to reconcile those differences when

---

*Ibid., p. 159f.

possible by a further refinement of theory or by noting differences in methodological procedure.

The propositional method is by far the most common approach to research aggregation. The academic discipline of psychology and related fields have a long and valuable tradition of review articles that has been the exemplar of aggregative reviews. Journals such as the *Psychological Bulletin* carry frequent articles that consider the research on a specified set of propositions. Special books or edited series such as *The Handbook of Social Psychology* will review the advances in specific fields that suggest one or another school of thought provides a more powerful explanation of some phenomena. Review is a continuing activity engaged in by large numbers of scholars because it is intrinsic to the cumulative development of a field. While most reviews deal with narrowly defined subjects, the scope can be quite broad. Perhaps the most sweeping study undertaken is *Human Behavior: An Inventory of Scientific Findings*, which sought "to present, as fully and as accurately as possble, what the behavioral sciences now know about the behavior of human beings."*

There is a wide range in the rigor employed in propositional reviews. The expository review generally takes the form of discussing one by one those studies the reviewer judges to be significant. The reviewer treats the well-known studies and those he feels that bear on some central question, summarizes their findings, and weaves an argument about what statements of relationship are supported by the evidence. The lay reader often cannot evaluate the reviewer's judgment about the relevance of the studies that have been omitted, nor can he judge the reviewer's objectivity in marshalling the evidence supporting and disconfirming the propositions being considered unless he is familiar with the literature.

Some propositional reviews are quite systematic. They consider a large number of studies and array them according to whether or not they support one or more propositions. One early but excellent example of a propositional review of a difficult literature is an analysis of the literature on science organization that sought determinants of the outcome

---

*Bernard Berelson and Gary A. Steiner, Harcourt, Brace & World, Inc., New York, 1964, p. 3.

"scientific accomplishment."* Anne Folger and Gerald Gordon found 88 studies dealing with how the leadership style, the nature of the research group, its context, and other factors affected scientific accomplishment. As Table 1 suggests, the number of studies that treated any

Table 1

SELECTED RELATIONSHIPS BETWEEN SCIENTIFIC ACCOMPLISHMENT AND ITS DETERMINANTS

Factor	Positive	Negative	Varying
Authoritarian leadership	--	7	--
Laissez-faire leadership	1	3	1
Participatory leadership	7	--	1
Multidisciplinary team	2	--	9
Academic institution	6	--	7
Industrial organization	1	2	8

one factor varied, as did the conclusions about the nature of the relationships. Although the approach makes it difficult to weigh and reconcile differences in the literature, one has a strong sense of what factors have been studied and the areas of consensus and disagreement about how they relate to scientific accomplishment.

Limitations of the Propositional Approach

When the propositions to be studied are carefully and explicitly defined and the literature is largely in agreement, the propositional review is an efficient and useful approach. If all of the ten studies on a given subject agree that there is a positive relationship between two variables, both the reviewer and the reader know what aggregative conclusion should be drawn. But what if the ten studies widely differ, with half concluding there is no relationship and half stating there is?

A forceful critique of this weakness of the propositional review is offered by Richard Light and Paul Smith. They characterize the method

*Anne Folger and Gerald Gordon, "Scientific Accomplishment and Social Organization: A Review of the Literature," *The American Behavioral Scientist*, Vol. 6, No. 4, December 1962, pp. 51-58.

of most reviews as listing the factors related to some key concept, excluding studies to create consistency, averaging some statistic across studies, or taking a vote so that each study of a relationship has one vote as to whether or not a relationship exists.* The difficulty is, of course, that each of these procedures has a serious weakness. Listing propositions can be a valuable guide for launching research, providing a systematic review of alternative causes or outcomes that might be taken into account. It simply avoids the task of aggregation. In some cases, listing is superior to excluding because at least the raw information is provided for the reader. Excluding, averaging, and voting approaches throw out information that might be useful in the analysis because they fail to deal with the central problem of interaction.

Interaction is a serious problem any time many studies come to different conclusions. To average out, to exclude, or to outvote some studies implies they are "wrong" and others are "correct." That assumption may or may not be true. It is also possible that the relationships under study may vary under different conditions. Thus one subset of studies may find that two variables, x and y, are not related because of the presence of a third variable that changes the nature of that relationship. If the majority of studies support the view that x and y are related, and a decision is made simply to reject the remainder of the literature, then it is quite possible that important information about interactive effects is being thrown out. There is the possibility of interaction due to the methods and approaches used in a literature as well as interaction among the factors being studied, for the existence or nonexistence of a relationship could be an artifact of the nature of the dominant methodology of a literature. Thus when the literature is found to disagree, the act of resolving the differences can be made more rigorous if there is a means to determine whether interactive effects are present.** The cluster method and the case survey method provide that capacity.

---

*Richard J. Light and Paul V. Smith, "Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies," *Harvard Educational Review*, Vol. 41, No. 4, November 1971, pp. 432-434.

**Using the propositional approach to test for interaction effects is possible, but it is a cumbersome process. All propositions must be

The Cluster Approach

There are two clear alternatives to the propositional approach to aggregation. Both are data approaches in that they use the research literature as a source of data rather than as a source of conclusions. Both were specifically developed to deal with the problems of interaction and contentious literatures. The cluster method of aggregating evidence to resolve differences is built upon the premise that "little headway can be made by pooling the words in the conclusions of a set of studies. Rather, progress will only come when we are able to pool the original data from the studies in a systematic manner."*

The central idea is to treat the data collected about individuals in different studies as having been drawn as "samples" from a common population. Through adaptation of the logic of cluster sampling techniques, fairly powerful statistical tests can be conducted to test whether the population under study or the relationships found studying that population differ from one research project to the next. If no differences are found, the clusters of data used by the different studies can be pooled, permitting analysis of the combined data. In such cases, one would expect the conclusions of the reanalysis to follow closely the original findings, but the outcomes would reach a higher level of statistical confidence. When the original data sets cannot be combined, the reviewer can try to isolate differences among studies that would account for different relationships and explore the nature of any interaction effects that had been uncovered.**

This approach is as limited as it is powerful. Because the reviewer is combining data in its original form, he is limited to those studies which used the same or highly comparable variables. In the area of education treated by Light and Smith, it is common to find studies that use the same measures of cognitive ability or school achievement. In some

---

tagged with information about some third variable, and separately aggregated. This step can only realistically be done for one or two variables without becoming so complex that it would be easier to use one of the alternative methods.

*Light and Smith, op. cit., p. 443.

**Ibid., pp. 445-464.

cases, when somewhat different measures are used, there exist accepted standards for calibrating one measure against the other. When one moves from the fields of education and psychology, however, studies using comparable data at the individual level cease to be very common. To use only those studies with comparable variables might thus entail using only a trivial proportion of the available studies in many fields. When the critical outcomes are organizational or programmatic variables and not individual characteristics, the cluster sampling logic is often not even relevant. And when the literature is not very quantitative, or did not use machine-readable data, the method cannot be applied. Thus there are limited fields, literatures, and review questions that can be addressed with the cluster method, but it remains that it is an extraordinarily powerful approach when it is appropriate.*

One of the strengths of the cluster method is derived from these limitations. The fact that the method is tied to studies with comparable variables means that often no new definitions and concepts are needed. The dependent or outcome variables can often be used exactly as they were and the subjectivity of the reviewer is not quite so likely to determine the outcome of the review. The same independent variables can sometimes be used, or flexibility can be introduced in the independent variables by testing for differences among subsets of clusters. The reviewer can proceed inductively and fish for groups of studies that can be pooled, and then determine their common characteristics. If, for example, studies of the impact of Follow-Through programs were found to predominate in one set of pooled data with common statistical characteristics, and studies of Montessori programs in another, then inferences could be made about the nature of the interactive effects caused by the two program types. Or one could begin deductively with studies grouped according to theoretical distinctions about the contexts of the various studies. These could be straightforward differences in the types of programs (e.g., Head Start,

*Nor should the practical difficulties in obtaining the original data be glossed over. Some scholars will view their data as proprietary and not release it or do so grudgingly. Data tapes that run on different computers with different formats must all be made compatible. These and other mundane difficulties have a tremendous capacity to consume time and energy.

Montessori), or they could be new variables with values assigned from a variety of sources (e.g., cost per pupil, or formal nature of the curriculum).

The Case Survey Approach

The third approach to aggregative research is the case survey. Another data approach, these methods put diverse case studies together in common conceptual terms. The cases can be clinical studies of individuals, administrative studies of organizations, anthropological reports on primitive societies, or any other set of descriptive analyses of a common social unit. To distill the lessons from these case experiences, the analyst prepares a set of questions to determine the presence and intensity of common characteristics, events, and outcomes contained in each of the case studies. The possible answers to the questions are carefully structured and defined so that the analyst, after reading the case materials, can readily determine the most appropriate response. The answers to these questions are determined in the same manner for each of the cases that have been selected for study. The results can then be put in a machine-readable form and analyzed.

Perhaps the best developed research of this type is the large body of research surrounding the Human Relations Area Files (HRAF). In an effort to bridge across the wealth of scattered anthropological field studies, a guide was published in 1938 for organizing and abstracting descriptive materials, and extensive indexing was carried out. These steps, under the Cross-Cultural Survey project, facilitated the systematic coding of cultural characteristics of the societies.* The presence and absence of these characteristics could then be correlated across societies to provide quantitative tests of hypotheses about cultural patterns. An extensive literature of comparative studies in anthropology has emerged whose scope is suggested by *A Cross-Cultural Summary*, which presented intercorrelations among 480 substantive variables and 56 methodological vari-

*George P. Murdock, "The Cross-Cultural Survey," *American Sociological Review*, Vol. 5, No. 3, June 1940, pp. 361-370; and "World Ethnographic Sample," *American Anthropologist*, New Series, Vol. 59, No. 4, August 1957, pp. 664-687.

ables for 400 societies.*

The vast nature of this undertaking and the desire to serve multiple theoretical concerns account for the massive impact of the Cross-Cultural Survey on the field of anthropology, but also create weaknesses. A check was run on the Human Relations Area Files by independently coding some of the original materials used for the HRAF. It was concluded that one "may expect that a careful study of the sources where the indexing was done by a person specifically trained for and sensitive to a particular research interest might find perhaps 25 percent to 50 percent more references than" the HRAF reported in its index.** The result is omitted case data. Moreover, the absence of a single, focused theoretical concern increases the difficulty of making borderline judgments about whether some social activity does or does not indicate the presence of a theoretical concept. Two respected and experienced anthropologists doing house-to-house research in Truk agreed in the abstract on the definition of terms, but could not agree on "the kind of residence each Trukese couple is in."† The same phenomenon can be seen differently from different theoretical perspectives, opening the way for subjectivity in the coding of the information in the case materials.

Whatever the disadvantages of the HRAF approach, its strengths make it well worth consideration as a model for research in other fields. In political science, the International Comparative Political Parties project has coded over 7,000 pages of materials on parties, but even though the scope of the effort has dropped from the parties of 90 nations down to 50, the project has consumed six years and is only now reaching substantive conclusions.†† Once completed, like the Cross-Cultural Survey,

---

*Robert B. Textor, HRAF Press, New Haven, 1967.

**Raoul Naroll and Donald Morrison, "Index to the Human Relations Area Files: Introduction," *Behavior Science Notes*, Vol. VII, 1972, p. 86.

†Ward H. Goodenough, *Description and Comparison in Cultural Anthropology*, Aldine Publishing Company, Chicago, 1970, p. 104.

††Kenneth Janda, "A Microfilm and Computer System for Analyzing Comparative Politics Literature" in George Gerbner, Ole R. Holsti, Klaus Krippendorff, William J. Paisley, and Philip J. Stone (eds.), *The Analysis of Communication Content*, John Wiley, New York, 1969, pp. 407-435. A prospective wide-range project is suggested by George D. Greenberg, Jeffrey A. Miller, Lawrence B. Mohr, and Bruce Vladeck, "Case Study Aggregation and Policy Theory," delivered at the American Political Science Association Meeting, New Orleans, 1973.

this project will make a major contribution to its field of research, both directly and through subsequent secondary studies. It remains, however, that the time and resources required for such undertakings mean that few can be carried out, and other, more narrowly focused aggregative approaches are also needed.

For aggregation to have the greatest value, it should also be an integral and continuing part of research. Thus it is possible to see the massive, multiple-purpose Cross-Cultural Survey as but one end of a continuum. Middle-level aggregations can slice across one or two major issues in a field to take stock of gaps in research, areas of consensus, and points of disagreement. At the other extreme, the individual scholar can review case studies for the relationships among two or three variables as a source of hypotheses in a new piece of original research.

An example of a middle-level, cross-case aggregation was developed over the summer of 1972 in a review of the literature on citizen organizations.* The aggregative research was started and completed in six months. The literature in question is highly diverse and contains many case studies of attempts to establish citizen influence in local community affairs and services. A preliminary consideration of the literature and of the issues of interest produced lists of factors considered important, and these factors were then expressed in a series of questions to be asked uniformly of the case studies. The questions, such as "Do the citizens have to sign off on applications for federal funds?" were framed with structured alternative answers. The analyst could answer "yes," "no," or "NW" if there was no way of determining the answers from the available materials. The resulting list of questions and alternative answers was then completed for each of 51 cases by a member of the reviewing team. He would read the article or book and fill in the checklist, making fairly straightforward judgments, e.g., the community is a rural area; the service is education; and the citizens have a role in the investigation of complaints. The checklist could then be put in machine-readable form.

---

*R. K. Yin, W. A. Lucas, P. L. Szanton, and J. A. Spindler, *Citizen Organizations: Increasing Client Control Over Services*, The Rand Corporation, R-1196-HEW, April 1973. Hereafter cited as *Citizen Organizations*.

The cases could then be used to study the correlation between various organizational characteristics and outcomes. Thus, when those cases where the citizen participation organization (CPO) did influence the complaint process were compared with those where they did not, striking differences were found in the relative success of the citizens in implementing their views about the program (see Table 2). The strength of the relationship had not been anticipated in the literature, and none of the 51 cases used as source material had emphasized the possible importance of the investigation of complaints. The aggregated results found a strong association missed by the individual case studies.

Table 2

Does the CPO have substantial influence in the investigation of complaints that individual citizens have about staff and program?

Program Impact	Responses	
	Yes (N=30)	No (N=13)
No or trivial implementation of citizen views	26%	77%
Significant or high implementation of citizen views	74%	23%
	100%	100%

It is important to emphasize that the checklist is not used as a questionnaire. For the aggregation of written materials, the citizens who took part in the case and the original researcher who reported that experience need not be contacted. The reviewer answers the questions based on the information in the written report, article, or book. It is the analyst's judgment that is put on the checklist. This is not to suggest that subjectivity of response has been eliminated, but judgments by trained analysts who have discussed and agreed on the meaning and nuance of the checklist questions, and who can discuss and clarify the intent of the questions with the research director, are qualitatively different from the views of diverse respondents.

Choosing an Approach

The choice of the method to be used to aggregate research must be determined by the nature of the literature being reviewed. If a literature is well defined, and if there is agreement in the literature, the propositional approach of the simplest sort may be all that is required. Particularly when specialists are writing for specialists, the expository propositional review is a satisfactory vehicle for scholarly debate because of the broader process in which it is embedded. Greater effort is required when disagreement is found in the literature, when methodological bias is suspected, or when substantive problems of interaction among the variables are likely. By and large, the cluster method can only be employed in limited contexts, but when it is applicable it is a powerful and important approach. The methodology is well developed by Light and Smith. What is needed there is experience in the practical problems of applying the approach.

The case survey method has corresponding but opposite strengths and weaknesses to the cluster approach. It has limited applicability to individual level data and can rarely be used when the central questions are about individual behavior or attitudes. Its strength is in its capacity to integrate the findings of diverse studies about organizations and programs. It is more flexible in that many different types of studies using different measurement techniques can be brought together, and new concepts can be developed and considered that none of the original research ever addressed. A good aggregation can be far more than a summation of what has already been said. This same flexibility also opens the way for its abuse.

Subjectivity and bias must be carefully guarded against in all three approaches, and there is considerable need for explicit and rigorous rules of procedure. It is to these rules that we now turn.

DECISION RULES FOR AGGREGATION*

These aggregative approaches can all be made more scientific (i.e.,

*The following discussion of decision rules draws heavily from the author's *The Case Survey Method: Aggregating Case Experience*, The Rand Corporation, R-1515-RC (forthcoming). A more extensive treatment of these considerations is found in that work.

more systematic, more rigorous, and less subjective) than research reviews usually are now. This is not to suggest that subjectivity can be eliminated. The reviewer's decision rules can and must, however, be made relatively explicit if aggregative policy reviews are to improve. Specifically, the reader must be provided explicit and consistent decision rules so that he is confident that another person with a different value position would reach the same conclusions if he used those same rules. The rules should be carefully developed and supported, but it is more important that the reader know what rules were employed than that he agrees with them. Known bias is always better than unknown bias.

Our first concern is those rules that should guide the search for relevant studies and the selection of those studies to be included. It is essential that the reviewer make explicit his criteria for exclusion and inclusion of studies so that the reader -- whether or not he agrees with the criteria -- can make his own judgment about the possibility that bias has entered the analysis. A more demanding requirement is that the reviewer should view the entire research literature as a non-random sample subject to bias, and take steps to determine the nature of that bias. As we shall see, these tasks can best be accomplished in a consistent and logical fashion if the purpose of the aggregation is well and clearly articulated. Indeed, the "sampling" concept of the research literature encourages the explicit definition of research goals, the universe of research under study, and the boundaries of the literature to be reviewed. But first let us consider the problems of exclusion and inclusion.

The Dilemma of Exclusion

Every research aggregation encounters the problem of studies that fail to meet acceptable standards of evidence. The reviewer finds methodological errors or an absence of scientific procedure that undermines his faith in the reliability of the evidence and the conclusions. To include such studies risks diluting or undercutting consistent patterns emerging in more reliable research, and so he may decide to omit them from his review. As soon as he has done that, however, those who question

his conclusions can point to all the evidence that was ignored, and suggest that the reviewer was guilty of sloppy scholarship, conscious bias, or both. The opposite of exclusion on methodological grounds is exhaustive inclusion, but that too has its own problems.

The difficulty of uncritical inclusion varies with the field, the quality of the literature, and the focus of the inquiry. A thorough search can be prohibitively expensive, but sampling procedures can deal with that difficulty. The major objection to exhaustive search and inclusion is that it leads to combining good research with bad, increasing the risk of unreliable results. Surely, a highly quantified study by a well-respected scholar is more valuable than an intuitive report by a person involved in the process being studied with a recognized axe to grind. Exclusion may or may not be better than undifferentiated inclusion, but the studies can be classified using the same criteria that would be required for exclusion. The early and creative attempt to do a research review of the determinants of scientific accomplishments mentioned above serves as an excellent example of some creative steps that can be taken in aggregative work, and the resulting strength of an approach that is inclusive but separates the studies according to their technical quality.

Anne Folger and Gerald Gordon aggregate the research on the organizational determinants of scientific accomplishment. They first make explicit the manner in which they define the literature.* A search of the Sociological and Psychological Abstracts and of 29 professional journals from 1950 to 1961 identified one set of reports on research productivity, and the citations in those sources led to the identification of further studies. In all, 88 studies were found. They briefly describe the nature of the studies, noting that 84 percent are by single authors. These probes of the literature's characteristics are tentative but insightful first steps toward a full identification of the sampling problem to be described below.

---

*Folger and Gordon, pp. 51-58.

The studies were then simply grouped into three categories according to the data they employed: "'hard' (systematic or structured studies), 'midway' (descriptive or unstructured observation), or 'soft' (speculative or personal experience)."* They then present the number and type of studies supporting the view that there is a positive, negative, or varying relationship between scientific accomplishment and other factors, such as the type of research leadership (see Table 3).

Table 3

SELECTED RELATIONSHIPS BETWEEN SOCIAL ORGANIZATION FACTORS AND SCIENTIFIC ACCOMPLISHMENT, BY OVERALL IMPRESSION OF DATA AND TREATMENT

Factor	Positive			Negative			Varying		
	Hard	Midway	Soft	Hard	Midway	Soft	Hard	Midway	Soft
Participatory leadership	1	3	3					1	
Amount or adequacy of funds		2	6					1	1
Long-term allocation of funds		2	3						
Adequacy of facilities		1	5			1		1	

The results suggest that participatory leadership is positively related to accomplishment. Had the authors combined the studies so that 7 studies of a quality not known to the reader were shown as supporting the existence of a positive relationship, the same conclusion would have been suggested, but the reader would not know that only one "hard" study supports that view. Had all the "soft" studies been excluded, evidence would have been lost that, while weak by itself, would add strength to the conclusion because it consistently supported other findings.**

This review has substantial weaknesses, but it stands out as a creative effort to aggregate a truly amorphous field in 1963. It avoids the problems of exclusion, and does not distort the results through naive

*Ibid., p. 54.

**The quality of the research can thus be used as a variable. See the "Data Quality Variables" discussion below.

inclusion. Its search method is explicit and plausibly exhaustive. What it and most other reviews fail to treat is the possibility that the literature itself is substantially biased.

The Sampling Concept of a Literature

There are different ways the literature available for analysis can be viewed. The first is to consider all available research on a subject to be the universe of phenomena under study. Each monograph, article, or book can be reviewed and conclusions reached about how the literature sums up. The choice of cases for research are made subjectively, however, and scholars hold common values and often choose to work in common settings. Thus the subjects chosen for study may well be a biased sample of the universe being studied.

Consider the commonly recognized problem of attitude research -- the overuse of students in introductory undergraduate psychology classes as subjects. The students are there, the cost of research is low, and careful laboratory experiments can be replicated. A vast preponderance of the research and the subsequent theory development in some areas of attitude research is consequently based on white, middle-class, college sophomores. What if one suspects that research on the poor, the very rich, the old, or those from minority cultures would lead to different conclusions, and there are no conclusions based on studies that test that suspicion? Conclusions based on "all available research" would then have to be carefully qualified as being based on biased research observations. This is not to say the conclusions are necessarily in error; only that of the universe of all po sible research studies, the available studies disproportionately represent observations of a particular kind. If the disproportion can be shown to be unrelated to other differences (i.e., if the processes of attitude change are not related to group differences) then the sampling bias can be set aside. But one must recognize when the "sampling" is nonrepresentative, and satisfy the reader that it does not affect the conclusions of the review.

Time is also a sampling parameter that can lead to bias. Case studies usually treat a flow of events. First, a program was started; then it won wider acceptance; then it became routinized in the broader

service activity it sought to supplement. Or, its leadership changed and its goals were altered. The effects and the inputs change over time. Thus the observations can often lead to very different conclusions depending on what time frame is chosen. If for some reason much of the literature is written at the same time, it is subject to the fads and enthusiasms that can sweep through the research community, the government, or society at large. Conclusions about citizen participation, for example, based only on organizations beginning in the early flush of the OEO's Community Action Programs that were written at that time might look quite different from conclusions written by the same observers about the same organizations during the mood of the early 1970s. Or, bias could be introduced because some stage in the development of programs is overrepresented. Conclusions on the effectiveness of Model Cities programs when they are all in the planning stage gives a different perspective than conclusions based on operating programs.

For some research purposes, it is necessary to use the same point of reference for all studies. That is not to say that all projects being reported on should necessarily be examined as they were, for instance, in the spring of 1968. Rather, it is to say that research on programs might, for example, look only at the six months before a program was initiated (if innovation was a concern) or at the second year of operation (if one sought to avoid the special effects of brand new programs). For other research purposes, the organizations or programs that have been selected should be studied at varied and representative stages of development.

By viewing the research literature as a set of observations and sampling a universe of all possible observations in both time and space, we can keep these types of problems in mind and check for the representativeness of the sample. Should bias be suspected, analysis can isolate the nature and degree of bias, and checks run to see if the bias is related to the conclusions of the review.

Research Goals and Sample Design

If the literature is seen as a sample of observations, what then is the universe of research observations? The fact that the sampling concept of the literature raises this question is one of its strengths:

A research aggregation, like any research, should be guided by explicit goals and a sense of theory. Criteria for what studies are to be excluded or included must be explicit and consistently applied, for as a practical matter there are many borderline cases that could be judged to be in or out based on substantive grounds. Is a study of a health information center at a rock festival to be included in an aggregation of research on decentralized information? If the review is focusing on the organizational determinants of continuing success, then it can be dropped. If the substantive question is how subcultural differences affect the efficient dissemination of information, then the study might be a critical observation of a variation not captured elsewhere. Vague research goals and fishing expeditions lend themselves to the scattered collection of studies, and ad hoc decisions on whether to include or exclude studies. The result of collecting studies for several undefined purposes opens the way to a set of observations that serve no single purpose well.

The decisions faced in formulating the design for a review of the literature on citizen participation illustrate some of the choices that need to be made.* The literature on the subject was quite extensive, and there were many studies of Model Cities and Community Action agencies in various locales. Had the purpose been an evaluation of citizen organizations in Model Cities programs, then the research universe would be all such programs, and the available research is a sample with unknown characteristics. The Model Cities studies would be listed and compared to the known universe to see if they were unrepresentative. The studies could then be weighted, so that the proportions of big city, regional, and other types of Model Cities were properly represented. Or one might do parallel analysis of big city and small city, Southern and non-Southern projects. These approaches have the advantage of using all the available studies, but one might also simply draw a quota sample from the literature. In any event, the universe, for the purpose of evaluating the organizational determinants of successful citizen participation in Model Cities or other programs, is the projects in those programs. Research reports on Model Cities are a nonrandom sample of observations of that universe.

---

*_Citizen Organizations_, op. cit.

Evaluation was not the purpose, however: The task was defined as an identification of the organizational determinants of successful participation across government programs, across locales, and across service areas. Since the number of observations (studies) of Model Cities and Community Action programs projects in the available literature outweighed other types of citizen organizations, to include all the literature without weighting would have led to results strongly biased by the experience of those two government programs. Thus the decision was made to limit studies in those areas with quotas, and to include as many other organizational variations as possible. The cases finally used were not chosen to be representative of the literature, nor were they representative of all past citizen organization experience. The quota sample sought the balanced collection of wide variation in organizational types, service areas, and locales because the universe of organization *types* was the phenomenon under investigation. Whatever the research purpose and sampling design chosen, a simple unweighted inclusion of all studies would have introduced bias and implied a study of the literature on citizen participation rather than a study of citizen organizations.

Unless one is concerned with the sociology of knowledge or the dynamics of research, and the conduct and process of research is the central concern, the universe of a research literature should never be taken uncritically as a set of observations. The literature should be searched in a systematic and exhaustive fashion, but the research thus located should be treated as a set of observations that could well be biased. Some forms of literature bias are unavoidable, if only those sources of difference that are associated with notoriety and the fact that the research itself was often an intervention in the social process. Whether or not these and other factors in fact are related to the conclusions of a review is not always clear, but the literature-as-nonrandom-sample concept should help keep the reviewer alert for possible bias. And certainly, if there is known bias, then the conclusions of a review must be tested for their sensitivity to the nature of the literature being included.

THEORY AND CONCEPT SPECIFICATION

Once the studies to be reviewed have been identified, one must then make explicit the decision rules that will guide aggregation. Where the rules for searching and selecting are generally applicable to all approaches, the need for rules for aggregation vary from one approach to the next. This discussion will emphasize the case survey method, but it applies to the propositional approach as well. Since the cluster method uses existing operational measures, concept specification has the standard problems and advantages of secondary analysis.

The greatest strengths and the fundamental weaknesses of the case survey method are the same: the almost infinite flexibility of the theories and concepts that can be studied. Those causes and outcomes central to any controversy will, of course, be considered, but there will always be an array of variables that may or may not be important to an understanding of the phenomenon under review. In practice, one cannot ask thousands upon thousands of questions of each case history, hoping to stumble across those mysterious factors that have a decisive influence. Some sense of theory is essential to bringing the inquiry into focus.

One thus begins with one or more theoretical models of how the phenomenon is best explained. It is not necessary and often constraining to limit oneself to a single theoretical model. Alternative models can be put forward, using different assumptions and different types of logical relationships. Indeed, one criterion for assessing the importance of a concept is to ask how many different theoretical explanations use it. If one or more crude models for ordering hypotheses about how the variables interrelate can be formulated, all the better. It may be sufficient, however, to lay out how classes of variables are logically related. Grouping variables may suggest important factors that are missing, and relating them may point up the importance of intervening or exogenous causes essential to a coherent theory. The first and central role of theory, and a working set of hypotheses, is thus an identification of the variables that should be included.

Level of Abstraction

The second role of theory, coupled with the purpose of the investigation, is to determine the level of abstraction that should be used in defining variables. To bridge across the studies, it is essential to find broad conceptual categories that allow different studies looking at slightly different phenomena to be pooled. Unless the literature in question is highly focused and addresses essentially the same propositions, to aggregate the studies requires aggregative concepts. These concepts must be broad enough to encompass the different concepts and measurement approaches in the individual studies. The definition of these concepts and the level of abstraction must be consistent with the purpose of the review and the theoretical questions behind that purpose. Moreover, it must be consistent with the size and nature of the literature being reviewed.

The need for aggregative concepts might be illustrated by a hypothetical example. Suppose a report argues that meeting the demands of the annual federal budgeting process prevents creative, ground-breaking research in federal labs; and an article shows the number of publications of biologists on one-year contracts is greater than academic biologists on grants which tend to run longer. The two conclusions can be left to stand separately, but the result of this approach would be long lists of somewhat related but different propositions, each supported only by one or two studies.*

To combine studies and aggregate them as evidence requires that broader concepts be developed to incorporate diverse findings. The definition of these aggregative concepts is the essential art and most subjective (and hence susceptible to bias) task in a research aggregation. Considerable precision is needed in defining each concept, and it aids the reader to provide examples of the lower-order terms, such as "number of publications," and "creative research" that are combined in the higher-order aggregative terms, "scientific accomplishment."

---

*For a listing of low-level proposition that gives a sense of multiple concept definitions even in a relatively focused literature, see Karen A. Heald and James K. Cooper, *An Annotated Bibliography on Rural Medical Care*, R-966-HEW, The Rand Corporation, April 1972, pp. 33-35.

The choice of the aggregative concept "scientific accomplishment" was in fact the key analytic decision in the Folger and Gordon propositional review of the organizational determinants of scientific accomplishment.* The practical problem is that given the number of studies (88) and their widely varied concerns, not many studies address comparable propositions. Thus the decision to go to a fairly high level of abstraction is forced if the number of studies on any given proposition is to be meaningful. Since the research purpose was to compare a range of fairly specific organizational factors, the choice was to increase the abstraction mostly of the outcome variable. The reviewers thus chose a very high level of abstraction for a single outcome variable, "scientific accomplishment," and aggregated propositions involving the alternative causes of that accomplishment.

There are, of course, alternatives. The review distinguished between two definitions of scientific accomplishment, and information could have been reported separately for scientific "productivity" and "innovation." This would not permit aggregation of the two conclusions hypothesized in our example, but aggregation could instead be expanded by moving to a higher level of abstraction among the independent variables. Thus "long-term allocation of funds" and the absence of "emphasis on adherence to deadlines" could be subsumed under "autonomy of scientist," and autonomy could then be separately related to productivity and innovation. The number of studies and the dispersion of the concerns in the literature force considerable abstracting; whether the reviewer abstracts the causes, the outcomes, or some combination of the two must be determined by the purpose of the review. But then each act of aggregating lower-order observations into higher-order conceptual categories must be made under explicit and consistent definitions, which are best developed within a common theoretical context.

Fact and Value

A common problem encountered in specifying concepts as they are

*Folger and Gordon, pp. 51-58.

found in an evaluation literature is that the conclusions are expressed in terms of value judgments. To say that a program or activity is "effective" or "successful" contains implicit normative considerations that are the domain of the decisionmaker. When identifying the factors related to positive and negative outcomes, the analyst doing the aggregation should treat the factual outcomes upon which the value judgments are based rather than the original research judgment about success. The reviewer may or may not then go on to express his views about whether the outcomes are good or bad, but he should strive to provide information that is as objective as possible about the outcomes under consideration.

Because the data approaches do not use the words or conclusions of the original research, they are thus better able to respond to this difficulty. If the literature contains value-laden research, the reviewer can create new aggregative concepts and set aside the conclusions actually drawn in the original research. In the citizen participation literature, there is common reference to success, failure, and effective participation, but those concepts have varied and even contradictory meaning. A program delayed or even blocked entirely by one citizen group would be called successful; a comparable delay in another case would be cited as an example of the dangers of citizen participation. Evidence of citizen militancy and overt conflict was seen as an advantage in one study; such conflict might have been judged as a disadvantage in another. The citizen participation literature involved so many concepts with strong normative overtones that a propositional integration required controversial value judgments that would almost certainly have undercut its objectivity. So instead, the checklist sought to ask questions of fact that could be answered by reading the reports, such as (1) whether policies favored by the citizens were implemented; and (2) whether policies they opposed had been blocked. Deciding that implementation or veto power was "good" or "bad" was a value judgment that could be kept separate from the data analysis.

Concept Reliability and Validity

One problem is expressed in terms of whether the checklist and the

machine-readable data it provides are a reliable reflection of the original case study. Often the case method will involve simple ordinal or nominal categories. The checklist will ask, for instance, whether there are specialized or functional committees in a citizen organization. If the original study includes that information, then it is straightforward to check "yes" or "no." One would expect the original researcher or another person carefully reading that same case study to come to the same conclusion. Many concepts are not so simple to treat, however. If one wants to know if those same citizens have "substantial influence" in investigating complaints, reasonable men may disagree on what the case study reveals. Detailed discussions among coders and explicit definition of terms are used to maintain consistency of judgment across cases, but one needs to know just how consistent the coders have been. One measure of reliability thus becomes the degree to which two different readers fill out a series of checklist questions in the same way for the same case. Standard coder reliability tests can be used for the entire checklist, or (if two or more readers have done large numbers of cases) for any given checklist item.

For every checklist item there must be a choice of saying that the case did not provide that information. Different researchers have different interests and concerns, and will not include description of varying aspects of the program or activity. Experience with the citizen organization review, however, found that it was often possible to make plausible inferences. But where in that event should the burden of proof reside? In a rigorous literature with rich information and consistent definitions, it is best not to infer. In a weak literature, to code the case as having insufficient data, except when the coder is quite sure, is to throw out a large proportion of the available information. To infer or (worse) to guess whenever there is some basis for selecting among the alternative answers to a checklist question blends a lot of low-confidence information with good data and greatly increases the possibility that coder bias will be introduced into the aggregation. As a simple expedient, the citizen organization study therefore introduced a level-of-confidence variable for each answer. If, for example, a yes-no response was appropriate, the coder could answer "yes," "no,"

or "not ascertainable," all to show high confidence. A second set of "yes" and "no" categories were choices if there was a reasonable basis to make an inference. Thus when a relationship was being tested across many cases, one had the capability to use all the information or to look at only the high-confidence data.

Reliability can be enhanced by careful instruction about the theoretical construct that is being pursued, by explicit definitions, and examples of how past ambiguities were resolved. In some cases, however, the greatest contribution to reliability was made by simply specifying the specific period to be used for each checklist. Case studies usually describe a varying program over time with varying effects. It is therefore essential to choose a consistent reference point in time for each case study.

Observer Reliability

Even if the coders agree on what the case study says, was the case study accurate in the first place? There are two ways of answering this question. The first takes advantage of whatever duplication might exist in the literature; the other requires taking the checklist to participants in the field knowledgeable about the cases.

Duplication is not uncommon in case study literatures. A particularly visible or accessible program will attract the attention of more than one observer, and two or more reports will appear in the literature. By having different coders initially complete a checklist for each study, however, one can determine how different observers see the same case, as reported on the checklist by different coders. The agreement between two such checklists is thus a measure of the combined observer and coder reliability. The duplicated cases are not a random sample of the available cases because they are by definition the more visible or more interesting. Therefore, the reviewer cannot argue that the observer-coder reliability of the duplicated cases applies to his entire set of cases. It is nonetheless a crude but valuable index. Since the universe consists of programs, not articles, the final aggregation should use only one combined checklist, so the two checklists must be reconciled and combined.

Whether or not duplicated case studies are available, the aggregative review can develop observer reliability measures by going to the field. An advantage of having a checklist in the form of questions is that those same questions can also be answered through field observation. By asking those questions of participants of the original program being studied, one can identify whether the report is biased toward one perspective or another. The analyst visits the site of the original study, and completes a set of checklists based on interviews with participants identified with different points of view. The checklists based on the views of these participants are then compared to the original checklist completed using the written case study.

An illustrative reliability check was run in the citizen organization research review. Four shortened checklists were completed, based on discussions in turn with (1) an elected consumer representative and (2) an appointed professional, both on the advisory board of a community health center; (3) the director of the center; and (4) the observer who wrote up the original case. A fifth checklist was completed by a coder working from the written materials.

Because of the importance of time as a reference point, error estimates will be somewhat high in this approach. The participants have been involved in an ongoing process, and it is hard to separate what was happening at an earlier time from more current events. The reviewer must ask for opinions about a time in the past -- that used in coding the case study onto the checklist -- even though memory data is always less than precise. But if the total agreement scores among the checklists are going to be low, the relative levels of agreement provide useful information about the possible bias in the case study. In our example (a current case with fewer memory bias problems), the levels of agreement with the checklist based on the written case are fairly good, although the consumer view is underrepresented. There is a common level of agreement among the informants, but it is interesting to note that the original observer disagrees more often with the consumer representative. Had more such reliability checks been possible and led to the same outcome, one would have an important measure of observer bias leading one to test the findings to determine whether they were artifacts of the perspectives of the original researchers. The absence of systematic bias, but a high

error rate, would suggest that there is considerable measurement error, and the reviewer should be reluctant to draw conclusions about whether two factors are related when no relationships were found in the data. Reliability checks of this sort thus test for observer bias in the original research, and are a conservative test for a maximum error rate (see Table 4).

Table 4

PERCENT AGREEMENT ON CHECKLISTS BY SOURCE

Source	Written[a]	Professional	Consumer	Director	Observer
Written case	--	91.3	62.5	87.0	87.5
Professional	91.3	--	42.9	60.6	66.7
Consumer	62.5	42.9	--	63.6	47.1
Director of center	87.0	60.6	63.6	--	62.5
Original observer	87.5	66.7	47.1	62.5	--

[a]The informants seemed likely to guess at times when a trained analyst would code that no information was available. Since the percent agreement score is based on matched responses when answers were shown on both checklists, the written code has a more reliable score.

Data Quality Variables

Another way of dealing with a weak literature is to create and use variables which are measures of research quality. The case survey method permits one to examine interactive effects that might be attributable to the nature of the method or observer. Thus the checklist should contain a series of judgments about the quality and type of research in order for the reviewer to determine whether those factors are related to the key findings of a review. In the review of the determinants of scientific accomplishment, it was found that autonomous scientists in academic settings are very successful. If there were data to show the degree to which these results were found only in reports by academics in academic journals, one could answer many important questions about observer bias. Had a case survey approach been used, and had the reviewer coded the institutional identity of the case observer and where the study appeared, he could analyze the data to determine whether those factors

made a difference. In a review of school features related to educational attainment, do quantitative and impressionistic studies lead to different conclusions? Again, the methodology employed could be coded as a variable and analyzed.

The idea of data quality variables, discussed by Raoul Naroll in the context of anthropological research,* has also been applied by Kenneth Janda in his aggregation of research on comparable political parties.** Janda codes the type of document, the language of the source, the national background of the author, his facility with the language of the nation being studied, a quantitative analysis score for the method used, and other variables that reflect on the reliability of the source. Thus his analysis based on the available literature can be exhaustive, including a broad range of sources; or he can be selectively exclusive and base his review only on quantitative studies by authors who are fluent in the language of the country and who use quantitative methods. And it permits one to determine whether the characteristics of the author or study are related to the findings. Using data quality variables avoids the dangers of exclusion without incurring the costs of excessive inclusion, and creates the capability to test for several forms of case bias.

Sampling Biases

Similarly, when the time and type of cases are coded as variables, the reviewer can examine the possibility that sampling bias has entered the research. If one fears that the studies done on citizen participation in the middle 1960s might be excessively optimistic, the checklist for those years can be deleted from the analysis if the year of the study has been coded. Or if Model Cities programs or programs in large cities are overrepresented, either deleting them from the analysis or using such

*Raoul Naroll, *Data Quality Control: A New Research Technique*, Free Press, Glencoe, Illinois, 1962.

**Kenneth Janda, "Data Quality Control and Library Research on Political Parties," *Special Problems of Comparative Method*, Chapter 46, pp. 962-974.

case "type" variables in multivariate analysis will cast light on how overrepresentation of types in the literature-as-nonrandom-sample might be affecting the findings.

CONCLUSION

It is remarkable that the same standards of scientific rigor applied to research projects are so rarely applied to the aggregation of research. Scientific research is expected to be systematic in coverage, rigorous in the specification of theory and measurement operations, and explicit in the form and results of analysis. Then a single scholar reads a series of such articles, argues through the results, and states his view of the summary judgments to be made. In light of the voluminous and fragmented nature of social research, the difficulty of obtaining studies, and the variety of research methods, more systematic approaches are now needed. Social research must develop a science of research aggregation.

If a research aggregation is to be more than a token effort to support intellectual and political positions already assumed, then it must convince the reader that the method of aggregation has no hidden bias. It is too much to expect a review to persuade everyone, but it will be vastly strengthened if it makes explicit the rules that were used to do the aggregation. At a minimum, the reviewer must delineate the body of literature he is considering, define his concepts carefully, and show the results of his review in an objective fashion to support whatever conclusions he might draw.

The cardinal rule of a good research aggregation is that the reviewer must provide sufficient evidence to enable the reader to make independent judgments about the conclusions. If he reader is left to accept the summary based on his faith in the authority of the reviewer and his relative agreement with the conclusions, then the reviewer has failed to conduct a scientific enterprise. The reviewer must apply as high or higher standards of rigorous inquiry to his aggregative work as he does to the separate studies he reviews.